NCCT Project Descriptions
BOSC Review, December 17-18 2007

Project Title: Information Management for Computational Toxicology

Lead Investigator: Richard Judson

Key Participants: NCCT: Ann Richard, Matt Martin, Imran Shah, Tom Knudsen, Fathi Elloumi, Keith Houck, David Dix: NHEERL: Stephen Edwards; Contractors (Lockheed-Martin): Marti Wolf, Tom Transue, Tommy Cathey, Amar Singh

1. NCCT Information Management Project: The NCCT Computational Toxicology Information Management effort is designed to support several of the strategic projects of the overall Computational Toxicology program. In particular, we are developing databases and tools to support the ToxCast program for screening and prioritization; the Virtual Liver and Developmental Toxicity projects in NCCT; and agency-wide efforts in genomics and chemical data management. Here we outline the most important current projects.

ACToR: (Aggregated Computational Toxicology Resource)[1, 2] is a database and set of software applications that bring into one central location many types and sources of data on environmental chemicals. Currently, the ACToR chemical database contains information on chemical structure, *in vitro* bioassays and *in vivo* toxicology assays derived from more than 150 sources including the EPA, CDC, FDA, NIH, state agencies, corresponding government agencies in Canada, Europe and Japan, universities, WHO and NGOs. In particular, the ACToR database includes all data from ToxRefDB, DSSTox (both described below) and ToxCast. ACToR uses a MySQL database and has a web-based front end for searching and browsing, and is being used within the Agency. We plan to make the system openly available through the internet in 2008. The full database will be made available to organizations that wish to install it locally. The short term goal of the ACToR project is to organize all of the data from the ToxCast program and to enable us to efficiently perform required analyses. The larger goal is to capture in one location all publicly available data on tens of thousands of environmental chemicals. This will allow us to easily identify data gaps for important classes of chemicals such as HPVs (High Production Volume) and pesticides. Preliminary results, which are in line with earlier analyses, show that for the vast majority of environmental chemicals, there is little or no publicly available toxicity data to inform risk assessment decisions. Within the next year, we will be using information in ACToR to help prioritize the selection of chemicals entering into Phase II of the ToxCast program.

ToxRefDB [3] is a database and user interface application that is being used to compile *in vivo* toxicology data on chemicals for the ToxCast program. Currently, the project is focused on capturing all relevant data from data evaluation records (DERs) on 280 food-use pesticides from the EPA Office of Pesticide Products (OPP). This set of chemicals is a subset of the initial 308 ToxCast Phase I chemicals. Currently, there are two versions of this database. All data entry uses an MS Access implementation, from which periodic copies are made to a MySQL version. Summary views of this database are then brought into the ACToR system.

DSSTox (Distributed Structure-Searchable Toxicity Database) Network [2, 4] is helping to build a public data foundation for improved structure-activity and predictive toxicology capabilities. In particular, the DSSTox group carefully curates chemical structure and related assay data for data sets of particular interest to environmental researchers. The DSSTox website provides a public forum for publishing downloadable, structure-searchable, standardized chemical structure files associated with toxicity data. The newly published DSSTox Structure-Browser enables structure-searching across or within individual DSSTox data files, bringing local structure-search capability to previously isolated EPA and National Toxicology Program (NTP) websites, including EPA's Integrated Risk Information System (IRIS), EPA's HPV Information System, and the NTP On-line Bioassay Database.

Genomics Data Management: Microarray-based genomics techniques are used widely in the NCCT, across ORD and in EPA program offices. Because these expensive-to-generate datasets are of potential use to scientists other than their initial creators, it was decided to create a single location that could house all genomics data generated at the EPA. The NCCT took the lead on this project and brought in ArrayTrack from the FDA to serve as this central genomics data repository. Currently, we are working with scientists in NCCT and NHEERL to load all currently existing microarray data from those organizations.

BDSM (Birth Defects System Manager) This resource provides a user-friendly and scientifically-required "integrative knowledge management system" for intelligent support to developmental biologists and toxicologists. At its core is a large, specialized database holding a comprehensive reference collection of gene-expression data for modeling animal development. Software applications and tools will be devised to link primary data with normal and abnormal developmental endpoints. A high-performance computing infrastructure and custom tools will mine these data to derive gene-expression signatures that can be directly compared between developing systems and across perturbations. This system is more extensively reviewed in its own session as part of the BOSC review.

Virtual Liver Pathway Database: This is a semantic repository used to manage biological pathway knowledge for the virtual liver program. It is intended to be a resource for internal and external collaborations on this project. All data either come from public sources or are EPA-generated data, which will be made public. This system is more extensively reviewed in its own session as part of the BOSC review.

ToxCast Data Management Workflow: The ToxCast program is generating large volumes of HTS, HCS (High-throughput and high-content screening) and genomics data on ~320 chemicals during Phase I [5]. To manage all of these data, we have developed a set of procedures for bringing in data from vendors, performing QC and backups and importing the data into ACToR. Once data are in ACToR, they can be aggregated across input data sets and exported for statistical analysis. The major application being used to QC the HTS/HCS data is GeneData Screener. This commercial product is widely used in the pharmaceutical industry and allows plate-by-plate QC, masking and normalization. ArrayTrack is being used for first pass QC and archiving of all of the ToxCast genomics data. Currently, we have data on >100 ToxCast assays entered into GeneData Screener.

ToxCast Data Analysis - ToxMiner: The analysis of the ToxCast data poses significant logistical and statistical issues. We have data from hundreds of chemicals and hundreds of assays, and there are potentially hundreds of endpoints to be predicted. In order to carry out and track all of the corresponding calculations, we are developing a set of software tools called ToxMiner. Currently, this is a set of R programs which can access the ACToR database to bring in data, and then carry out calculations based on requests submitted to a set of workflow database tables. The results are again stored in the database for further analysis. The basic statistical approach is to apply one or more machine learning algorithms to find classifiers that link *in vitro* ToxCast data to *in vivo* phenotypes captured in ToxRefDB. We have been investigating the performance of a number of learning algorithms using simulated data [6]. The basic conclusions are that several learning algorithms work relatively well (support vector machines (SVM), artificial neural nets and linear discriminant analysis are the most promising), although all suffer drops in performances as the amount of experimental noise and the number of irrelevant assays increase.

2. What is the EPA context for the project?

One of the most pressing problems faced by the EPA and other environmental regulatory agencies is the huge number of unique chemicals found in the environment, coupled with the fact that very little is known about the vast majority of these chemicals. In The US, there are ~10,000 chemicals that are either pesticide ingredients, are HPVs or are otherwise identified as posing a significant exposure risk. There are an additional ~20,000 chemicals in the so-called MPV (medium production volume) class which pose some exposure risk. The total number of unique chemicals in the environment is somewhere in the range of 80,000-100,000. This should be contrasted with the number of chemicals for which the complete battery of gold-standard animal-based toxicology studies have been performed, which is probably less than 2,000. It would not be possible to assess in a timely fashion the majority of even the 10,000 most prevalent chemicals using the gold-standard complete battery of tests because of the cost (~10M-$20M/chemical) and the number of animals which would be required.

This disparity drives the need for so-called screening and prioritization projects, of which ToxCast is the leading example in the EPA. The goal of these projects is to develop *in vitro* or *in silico* “signatures” that are predictive of particular toxicities and also much less expensive to produce than the corresponding animal data. Chemicals which show particular *in vitro* signatures would then be prioritized for more detailed testing, including relevant *in vivo* studies. ACToR and ToxRefDB are directly helping these efforts by compiling all of the publicly existing chemical toxicity data in one place. This allows a thorough assessment of the current state of knowledge and the corresponding data gaps. Data in these databases will be used for training and validation of ToxCast signatures. Because these resources will be made publicly available, they will also help other screening and prioritization efforts both within and external to the EPA.

Moving to an *in vit*ro / *in silico* approach to chemical toxicity evaluation has a second advantage beyond cost. These new approaches are inherently probing and analyzing toxic effects at a mechanistic level and should therefore provide unique insights for carrying out risk assessments on particular chemicals. The Virtual Liver and Birth Defects project are building detailed mechanistic models of particular toxicity processes or pathways. The ToxCast program, by coupling multiple biochemical, cell-based, genomics and *in vivo* assays for the same set of chemicals, will provide evidence for detailed mechanisms / modes of action for a variety of toxic chemicals.

3. What are the strategic directions and science challenges?

All of these projects support NCCT screening and prioritization efforts. ToxCast is the overall program for generating screening data and ToxRefDB, ACToR and DSSTox projects provide data capture, management, QC and analysis support. The Virtual Liver and BDSM will help provide mechanistic understanding of the ToxCast signatures, in addition to meeting their own goals. DSSTox and ACToR provide broader resources for information on environmental chemicals. These resources will be of use to scientists and regulators outside of the NCCT for performing modeling studies and predicting chemical toxicity.

The NCCT genomics data management project is providing infrastructure for scientists and regulators throughout the agency to manage and archive genomics data in a safe and secure environment. This allows sharing of genomics data between groups that will facilitate meta-analyses and support cross-organization collaborations.

We face several significant challenges with carrying out these projects.

Information heterogeneity / Information Integration – Our ultimate goal is to build tools for understanding and predicting chemical toxicity. The development of these tools will rely on many types of data from many sources in many formats. ACToR is pulling in data from over 150 different sources. Some of the data are tabular, some are relational, but much of the data are embedded in text reports. This is also true for ToxRefDB, BDSM, Virtual Liver and DSSTox. In order to be aggregated in ways that are useful for computation and modeling, the information must be extracted and formatted. This can always be done manually, but this is too labor intensive in many cases, so we will need to develop automated tools to do the majority of information extraction and integration from diverse data sources.

Another integration challenge is that information can have a complex, hierarchical structure. We need to design data models that are general enough to be applied to multiple data sets yet specific enough to make it practical to map source data into the aggregated data repository. ToxRefDB has made use of insights from other toxicology informatics projects to develop a robust relational schema for capturing *in vivo* toxicity data.

Another challenge worth mentioning involves text mining. There is a significant body of literature on toxicity studies for chemicals of interest, but manual literature mining will never allow us to make significant progress in evaluating toxicity for thousands of chemicals. The alternative is to use automated text mining methods at least as a first filter of the literature to discover interesting linkages between chemicals and toxicity data. We are carrying out a pilot project to mine the OPP DER records that are the source of the ToxRefDB data. This will provide an estimate of the sensitivity and specificity of these automated approaches for capturing useful information.

Quality Control / Information Freshness – Two other related challenges are that of assessing the quality of data in our databases and insuring that the data are fresh or up-to-date. For data sources for which downloadable, tabular data are available, we do not do further quality control, but instead typically accept data as is. For other sources, we have to automatically or manually extract data, tabularize it, and then load into a database. These data all need to go through some level of quality control, but for most cases, a 100% manual QC is too expensive. DSSTox and ToxRefDB are exceptions to this, and do go through labor intensive QC processes. For the rest of the data in ACToR and other sources, automated QC processes need to be devised. Most of the sources from which we are extracting data are not static, so it is important to have a process to keep our version up-to-date. For the ACToR project, this can be as simple as downloading the most recent version of an included database on a monthly basis and refreshing the local database. This is the case with PubChem, whose developers intend other groups to download and make local use of the data. For most other sources, though, the data is simply a set of web pages, so we need to automate the process of looking at the web resource on a periodic basis to detect and bring in changed content.

Data to Knowledge – Perhaps the greatest challenge in the information management arena is that of converting data or information into knowledge. ToxCast and the Virtual Liver project are good illustrations of this issue. The first phase of ToxCast analysis will find statistical correlations between the results of a set of in vitro assays and whole animal toxicity endpoints. In a sense, these correlations or associations are simply hypotheses about linkages between direct molecular processes (for instance receptor binding) and more complex biological processes. However, only by validating these initial findings through more detailed mechanistic studies and model building (as in the Virtual Liver) can we have sufficient confidence in the linkage to use it for predictive risk characterization.

Barriers to Success – The ToxCast program has several potential barriers to success that relate to information management. The first is the possibility that the source of *in vivo* toxicology data in ToxRefDB is not of sufficiently high quality or that there is not enough data to allow us to find “true” *in vitro* to *in vivo* associations. An alternative way to state this is that the mapping from initial molecular interactions to final whole-animal toxicity may be so complex that we do not have sufficient power in the current study to find true relationships. This is not a fundamental limitation, but is instead a statement about available resources.

4. What are the short-term (1-2 year) and long-term (3-5 year) goals?

Short term goals (1-2 years)

1. ToxCast IM Goals
    - Complete implementation of ToxCast analysis tools for classifier discovery
    - Complete implementation of ToxCast workflow and data QC pipeline.
    - Use ACToR and ToxRefDB to capture phenotype data for candidate compounds for ToxCast Phase II (ToxCast, ToxRefDB, ACToR)
    - Use DSSTox to provide chemical structure annotation and QC of chemical information
    - Complete first phase of ToxCast data QC and analysis and produce a set of candidate signatures for use in ToxCast Phase II.
2. ACToR Goals
    - Complete first phase of data entry for ACToR
    - Make ACToR web site fully functional and available on the internet
    - Complete analysis of data landscape for target environmental chemicals
    - Add workflow capabilities to facilitate construction of ToxCast Phase II toxicology data sets

3. ToxRefDB Goals
    - Complete entry of all current DER records
    - Develop approach to appropriately capture non-DER data in ToxRefDB
    - Capture data for all ToxCast Phase I and candidate Phase II chemicals
    - Make ToxRefDB data publicly available through the ACToR web site.
4. DSSTox Goals:
    - Apply strict DSSTox Chemical Information Quality Review Procedures to the annotation of chemical structures and chemical information associated with toxicity data and environmental chemicals across EPA programs;
    - Coordinate with outside public efforts to encourage chemical structure annotation, data standardization, and open public access to toxicity data files and public databases
    - Provide full, open access to toxicity data files for chemical structure-analog searching, and for facilitating development of improved models for predicting toxicity based on chemical structure;
    - Use DSSTox to link users to large public resources, and to link those large public resources to source data websites, such as in EPA and the NTP.
5. Genomics Data Management
    - Import all existing ORD genomics data into ArrayTrack
    - Develop procedures to bring in all new genomics data

<u>Long Term Goals (3-5 years)</u>

1. Develop overall screening and prioritization framework for ToxCast program, combining data in ToxRefDB and ACToR with external tools, and adding domains such as exposure risk and economic impacts. This will require close partnerships with relevant Program Offices.
2. Develop formal evaluation and pre-validation approach for ToxCast Phase I and II predictive signatures.
3. Upgrade informatics infrastructure to more rigorous statistical procedures and better QC and archiving of data.
4. Merge information gained from mechanism-based projects (Virtual Liver, BDSM) with that from ToxCast to produce approach to classify chemicals based on detailed mechanism and mode of action, rather than on more statistically-based ToxCast signatures.
5. Expand the BDSM prototype to other developing target organ systems and integrate with NCCT/ORD research programs for biological modeling and regulatory analysis, including ToxCast and the Virtual Liver.
6. DSSTox: Expand the range of published DSSTox data files to include ecotoxicology, public genomics inventories, and specialty sets amenable to modeling across domains of toxicology (e.g., immunotoxicology, neurotoxicology, skin sensitization, etc) Create seamless interfaces between DSSTox structure-browser and outside resources such as ACToR and PubChem to optimize functionality of each. Use DSSTox to interface ToxRefDB and ToxCast with Structure-Activity Relationship (SAR) modeling community by separate, documented publication of summary activity data files amenable to modeling

5. What other components of EPA or outside organizations are involved?

EPA Collaborators Several of these projects are being performed in collaboration with Program Offices. The development of ToxRefDB is a joint project of Office of Pesticide Programs (OPP) and NCCT. ACToR is supporting a collaborative project with OPP, Office of Pollution Prevention and Toxics (OPPT) and Office of Water (OW) to understand the data landscape for environmental chemicals. The genomics data management project is a collaboration between ORD and (Office of the Science Advisor) OSA and is intended to help provide the genomics data management and data analysis tools for OPP, OPPT and OW. The analysis of the ToxCast Phase I data will be carried out jointly by staff in NCCT, OPPTS and outside collaborators.

External Collaborators:

ACToR Data Sources: ACToR is extracting data from many sources outside the EPA, and we are being actively aided by researchers at NIH (PubChem project and National Cancer institute), and Health Canada. We are in initial discussion to share data with the OECD QSAR working group as part of the QSAR Toolbox project.

ToxRefDB Collaborations: ToxRefDB has benefitted from information sharing with other parts of ORD, OPP and OPPT, NIEHS, the ToxML working group, the ILSI-HESI developmental toxicology database project, and the RepDose team at the Fraunhofer Institute (Germany). Along with OPP, we are planning to share the ToxRefDB database with regulators in Germany, the Netherlands and Canada, as well as with other ToxCast collaborators through formal arrangements (e.g., Material Transfer Agreements).

Genomics Collaborations: NCCT researchers have been long-term participants in the Microarray Quality Control (MAQC) project started by the FDA. We are currently working with a 29 groups around the world to evaluate methods for evaluating large toxicogenomics datasets.

EPA STAR Centers: The EPA is funding 2 large center grants in the area of bioinformatics and toxico-informatics. These are in North Carolina (University of North Carolina, Chapel Hill) and New Jersey (University of Medicine and Dentistry of New jersey and Princeton University). We have active collaborations with both of these centers in the areas of genomics analysis, chemoinformatics and pathway modeling.

ToxCast Analysis Partners: We are in discussions with several companies about jointly analyzing ToxCast Phase I data. These companies will provide the EPA access to software tools and expertise, in exchange for early access to the ToxCast data. Similar arrangements are being discussed with academic groups.

DSSTox Collaborators: We are working with several groups inside and external to the EPA to publish structure-annotated toxicity data files, to provide linkages to external data and structure searching capabilities. These include HPV information system (HPVIS) from OPPT, IRIS and the NTP, and public resources such as PubChem and ChemSpider.

BDSM and Virtual Liver collaborations will be described in their respective sections.

6. How is data management being achieved?

Software: For many of the projects described here, we have tried as much as possible to use open source software. ACToR and the Virtual Liver database are built using MySQL. All of the ToxCast data will go into ACToR, so will be available as a set of flat files or as a complete database dump. The web interface for ACToR is written in Java. Database loading and QC tools are written in Perl and Java. Chemical structure tools are primarily built using the open source CDK library. We are making certain improvements to this library and sharing them with the core developers. The first choice for the development of ToxCast analysis tool is R. ToxMiner is currently a combination of R, Perl and Java.

ToxRefDB is built using MS Access, principally for ease of use by individuals doing data entry. This has prevented us having to deal with intractable network security issues. However, the final version to be used for analysis and integration with ToxCast is implemented in MySQL and is tightly linked with the other ACToR databases.

The primary output of DSSTox is a series of ASCII structure files which are freely accessible from the NCCT web site. The DSSTox Structure-Browser is built on publicly accessible tools and open source software.

ArrayTrack is a “free” application for managing microarray data. Currently, however, the backend is Oracle Enterprise addition. The FDA and their collaborators at UMDNJ are developing a version that will use one of the open source database applications, and we will switch to that version once it becomes available.

BDSM uses Oracle Standard Edition as the backend, but all other software is developed in-house and could be made available to collaborators. This software is written using a combination of Perl, Java and R.

Hardware: In addition to up-to-date desktop machines, the NCCT staff has access to several Dell dual processor dual / core servers, several high end desk side servers and the EPA National Computer Center’s (NCC) SGI Altix Supercomputer. During FY08, we will acquire a high-end computer system for chemical structure management. This will include multiple additional servers in addition to specialized commercial software.

7. What are appropriate measures of success?

- ACToR:
    1. Completion of input of all identified data sets relevant to environmental chemicals.
    2. Capture of all ToxCast and ToxRefDB data
    3. Availability of the system on the Internet
    4. Use of ACToR by program offices
    5. Use of data by wider community, including OECD and Non-US regulatory agencies
    6. Collaborative use by academic groups to mine this large data collection.
    7. Use to develop ToxCast Phase II training and validation set.

- ToxRefDB
    1. Completion of annotation of all ToxCast Phase I chemicals

2. Completion of QC of all data
3. Use by ToxCast data analysis project
4. Use by OPP data analysis projects
5. Use by outside groups including other regulatory agencies, OECD and scientific collaborators.
6. Availability of the data on the Internet as part of the ACToR system.

- DSSTox
  1. Adoption of DSSTox standards and QA procedures, and use of the structure browser by outside groups
  2. Incorporation of DSSTox data files into outside public resources such as PubChem, ChemSpider, and the OECD Toolbox.
  3. Expanded collaborations and publication of new DSSTox data files.

- Genomics Data Management:
  1. Completion of input of the majority of ORD genomics data
  2. Capture of all ToxCast Phase I genomics data
  3. Use of the system for analysis of appropriate data sets (including ToxCast Phase I genomics)
  4. Completion of MAQC II data analysis project

- ToxCast Data Management Workflow:
  1. Capture and QC of all Phase I ToxCast HTS/HCS data
  2. Transfer of all data into ACToR
  3. Organization of ToxCast Phase I data that will facilitate ToxCast data analysis

- ToxCast Data Analysis:
  1. Completion of descriptive statistical analysis of all ToxCast Phase I data by the end of May 2008
  2. Significant progress on development of candidate predictive signatures from ToxCast Phase I by the end of summer 2008

8. References

1. Judson RS, Richard AM, Dix DJ, Houck K, Elloumi F, Martin MT, Cathey T, Transue TR, Spencer R, Wolf MA: **ACToR – Aggregated Computational Toxicology Resource.** *Toxicology and Applied Pharmacology* 2008, **Submitted**.
2. Richard A, Judson R, Yang C: **Toxicity Data Informatics: Supporting a New Paradigm for Toxicity Prediction.** *Toxicology Mechanisms and Methods* 2008, **In Press**.
3. Martin MT, Houck KA, McLaurin K, Richard AM, Dix DJ: **Linking Regulatory Toxicological Information on Environmental Chemicals with High-Throughput Screening (HTS) and Genomic Data.** *The Toxicologist CD- An official Journal of the Society of Toxicology* 2007, **96:**219-220.
4. **DSSTox EPA Integrated Risk Information System (IRIS) Toxicity Review Data: SDF File and Documentation** [www.epa.gov/ncct/dsstox/]
5. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ: **The ToxCast program for prioritizing toxicity testing of environmental chemicals.** *Toxicol Sci* 2007, **95:**5-12.
6. Judson R, Elloumi F, Setzer RW, Li Z, Shah I: **A Comparison of Machine Learning Algorithms for Chemical Toxicity Classification Using a Simulated Multi-Scale Data Model.** *Manuscript in Preparation* 2008.